IMPACT: International Journal of Research in Business Management (IMPACT: IJRBM)
ISSN (P): 2347-4572; ISSN (E): 2321-886X
Vol. 5, Issue 7, Jul 2017, 25-30

ROBUST ESTIMATION OF LOCATION AND SCATTER, BASED ON SUBSET SELECTION PROCEDURE WITH APPLICATION IN DISCRIMINANT ANALYSIS

K. MAHESH

Department of Statistics, Government Arts College, Udumalpet, Tamil Nadu, India

ABSTRACT

Multivariate statistical techniques are most widely used in basic sciences research. These techniques, such as classification and data reduction methods, mainly rely on the two estimators, location and scatter. The sample mean and covariance matrix is calmer, money is used as estimator. But, they are extremely sensitive to elliptical distribution with heavy tails. In this context, many robust alternatives are entrenched. The accuracy of these robust estimators mainly based on ***'h'*** data points out of ***n***. This paper suggests a robust procedure of selecting ***'h'*** data points, in order to get closely three estimates. It demonstrates the efficiency of the proposed procedure, by applying it in classification method under a real environment.

KEYWORDS: Location and Scatter-Robust Estimators - Discriminant Analysis.

INTRODUCTION

In statistics, the location and scatter parameters play a vital role in all multivariate statistical techniques. Conventional estimate of location and scatter are the sample mean and covariance matrix, which are very sensitive to outlier. In this context, many robust alternatives are entrenched during the past decades. The basic multivariate robust estimators are minimum volume ellipsoid (MVE) and minimum. Covariance determinant (MCD) was proposed by Rousseau (1985). These methods compute the parameters by finding the h data points, out of n points in the given data set. Hence, the estimated parameters rely only on the h data points. It is necessary to select suitable h data points in order to compute the parameters. There is no computationally reasonable group wise best point selection procedure of calculating MCD and MVE.

The MVE and MCD of a given data set is determined by a subset to the constraint that the ellipsoid that covers the point has minimum volume and minimum covariance determinant among all constructed, using *my* points (Rousseeuw (1985), Hawkins (1993) and Woodruff and Rocke (1993)). The size of the subset is a function of the number of the data points' n and the dimensionality p, and is chosen to give an estimate with a breakdown point of 50%.

In this context, it is proposed the best point selection (subset selection) procedure, based on MVE and MCD. The description classical estimator, maximum likelihood estimator (MLE), robust estimators MVE and MCD along with Feasible Solution Algorithm (FSA) are provided in the section 2. The proposed procedure and its computation steps are presented in the section 3. The assessment of the proposed procedure over the other procedures has been studied in the context of real environments, and the results are summarized in the section 4. Summary of the findings is presented in the last section.

Classical and Robust Estimators

The theoretical aspects of the classical and robust estimators, MLE, MVE and MCD along with FSA are briefly furnished in this section.

Maximum Likelihood Estimator

The principle of maximum likelihood estimation was originally developed by Professor R.A Fisher in 1920. The standard estimates are the maximum likelihood estimates or their unbiased variance. The sample mean vectors for the provision of the estimates of $\hat{\mu}_i$ are given by

$$\hat{\mu}_i = \bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij} / n_i \ , \quad (\text{i=1,…,g}) \tag{1}$$

For the hetroscedastic case, each Σ_i is estimated by its training sample analogue usually after correction for bias, to give

$$\hat{\Sigma}_i = S_i = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{ij})(Y_{ij} - \bar{Y}_{ij})' / (n_i - 1). \ (\text{I=1,…, g}). \tag{2}$$

For the homoscedastic discriminant analysis model, the standard estimate of the common covariance matrix $\sum$ is pooled within – group sample covariance matrix

$$\hat{\Sigma} = S = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}i.)(Y_{ij} - \bar{Y}i.)' / (n - g), \tag{3}$$

Where $n = \sum_{i=1}^{g} n_i$ is the total sample size across groups?

Minimum Volume Ellipsoid (MVE)

The minimum volume ellipsoid estimator was proposed by Rousseau (1985). The MVE estimator is a very powerful procedure for estimating the location and scatter. It is known that it has a breakdown point that approaches 50% as the number of points in the data set increases. This is the maximum possible breakdown point, and it means that approximately half of the data can be arbitrarily contaminated without affecting estimate. The computational steps are as follows:

$$\text{Let, } (x-c)^T \Gamma^{-1} (x-c) = p, \tag{4}$$

Where c and Γ are location vector and scatter matrix respectively and p is the dimension of the data. The location vector is the weighted mean calculated as

$$c = \sum_{i=1}^{h} w_i x_i^*, \tag{5}$$

And, the covariance or scatter matrix is

$$\Gamma = \sum_{i=1}^{h} w_i (x_i^* - c)(x_i^* - c)^T, \quad \textbf{(6)}$$

Where x_i^* is a cothector denoting the i th observation ofthe subset of points, w_i is the weight for the i th observation, and $h = [(n + p + 1)/2]$, ([.] denotes the greatest integer function). The volume of the covering ellipsoid will be proportional to the determinant of Γ. It is evident from these equations that to find MVE one must determine which he points should be covered and the corresponding weights to ensure coverage of the points.

Minimum Covariance Determinant (MCD) Estimators

Rousseau (1985) introduced the minimum covariance determinant estimator (MCD) to estimate the mean vector and covariance matrix along with detection of outliers in multidimensional data. The multivariate location and diffusion estimation in high breakdown principles are based on the determinant of the covariance matrix. The covariance matrix *n x p* is positive semi-definite matrix, *p* eigen values are positive, the determinant of the covariance matrix equals the product of the eigenvalues. Thus, a small value in the determinant reflects some linear patterns in the data. Consider all C_h^n subsets, and compute the determinant of the covariance matrix for each subset. The subset with smallest determinant is used to calculate the usual $p\,x1$ mean vector, and corresponding pxp covariance matrix, these estimators are called minimum covariance determinant estimators.

Hawkins (1994) established a feasible solution algorithm (FSA) for MCD approach and its brief descriptions is as follows. The FSA is to reduce the determinant of the matrix further by pair wise case swap. The trial partition of the cases into trimmed and retained cases. It involves evaluating pair wise exchanges between retained case and a trimmed case. The pairwise exchange will lead to a reduction in the covariance determinant. It is clear from the definition that the MCD estimator for μ and Σ is the sample mean vector and covariance matrix of a subset of size $n-h$, the determinant of $\hat{\Sigma}$ cannot be decreased by any case wise exchange exchanging one of the trimmed cases for one of the retained cases.

Best Points Selection Procedure (BPS)

MVE and MCD procedure of interest is to find the points out of n, the total cases of all the groups considered. The main limitation of these methods is to extract the large number of subsets of h points out of total n points. For the voluminous of given data, computational point of view it is very difficult to select h points out of n by satisfying the criteria of minimum volume and minimum determinant. The proposed Best Points Selection (BPS) algorithm based on MVE and MCD is to find the weighted mean and weighted covariance matrix, based on the selected h points. The proposed BPS procedure is summarized given below.

First, select the h points in the each group separately by either MVE or MCD procedure and then compute location and scatter for the combined all the h points together by giving equal weights. Next, compute the Mahalanobis distance for all observations based on the computed weighted location and scatter. Arrange the distances, the first p+j (j=1,2,…, h-p) distances are selected and their corresponding sample units are used to compute the next **Error! Bookmark not defined.**and s. Repeat the same procedure until h data points are selected. Then compute location and scatter for the

selected h points. The general description of the computational algorithm is as follows.

Step 1: Let $X=(x_1,x_2,...,x_m)$ be a set of m points in $\Re^P$, let h be a natural number such that $m/2 \le h = \frac{(g_1+g_2+...+g_n)}{([m+p+1]/2)} < m$.

Step 2: The $p+1$ data points $\{x_1,x_2,...,x_{p+1}\}$, that satisfy the two optimality criteria were selected and use to obtain the location and scatter matrix

$$\bar{x}=\frac{1}{p+1}\sum_{i=1}^{N} w_i x_i \text{, and } s_{jk}=\frac{1}{p+1}\frac{\sum_{i=1}^{N} w_i}{\left(\sum_{i=1}^{N} w_i\right)-\sum_{i=1}^{N} w_i^2}\sum_{i=1}^{N} w_i(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$$

Step 3: Compute the Mahalanobis distance for all the observations using $\bar{x}$ and S, as $d_i^2=(x-\bar{x})S^{-1}(x-\bar{x})^T$

Step 4: The $d_i^2(i=1,2,...,k)$ is arranged in order of magnitude from the least to the highest. The first p+j (j=2, 3, 4… h-p) distances are selected and their corresponding sample units are used to compute the next **Error! Bookmark not defined.**and s as follows

$$\bar{x}=\frac{1}{p+j}\sum_{i=1}^{N} w_i x_i \text{ And } s_{jk}=\frac{1}{p+j}\frac{\sum_{i=1}^{N} w_i}{\left(\sum_{i=1}^{N} w_i\right)-\sum_{i=1}^{N} w_i^2}\sum_{i=1}^{N} w_i(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$$

The new set of $\bar{x}$ and s are then used to obtain the Mahalanobis distances for all the observations.

Step 5: Step 3 and 4 are repeated until the number of units selected is $h=\left(\frac{(g_1+g_2+...+g_n}{([m+p+1]/2)}\right)$.

Finally, the location and scatter based on the BPS algorithm can be computed as

$$\bar{x}_{(BPS)}=\frac{1}{h}\sum_{i=1}^{N} w_i x_i \text{ , and } s_{jk(BPS)}=\frac{1}{h}\frac{\sum_{i=1}^{N} w_i}{\left(\sum_{i=1}^{N} w_i\right)-\sum_{i=1}^{N} w_i^2}\sum_{i=1}^{N} w_i(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$$

Experimental Results

This section presents the summary of the discriminant analysis results, specifically confusion matrix, apparent error rate and discriminant coefficients, which were obtained based on the classical, robust procedures along with the proposed procedure. First, compute the mean vector and covariance matrix, and then performed the discriminant analysis for the given training data set. Secondly, the same discriminant function used for the validation data to validate the function.

The iris data set was considered for this experiment. Also, the data set was divided into two categories, the 60% of the data were considered as training data and the remaining 40% considered as validation data. The results obtained for

the training data under the various procedures are summarized in the table 1.

Table 1: Results of Discrimination Analysis under the Various Procedures (Training Data)

Methods	MLE	MVE	MCD	FSA	BPS_MVE	BPS_MCD
Discriminant Function	$\begin{pmatrix} LD1 & LD2 \\ 0.58 & -1.26 \\ 1.90 & 3.07 \\ -1.77 & 0.32 \\ -3.52 & 1.41 \end{pmatrix}$	$\begin{pmatrix} LD1 & LD2 \\ 0.12 & -1.70 \\ -1.90 & 3.40 \\ -1.07 & 0.17 \\ -6.00 & 2.26 \end{pmatrix}$	$\begin{pmatrix} LD1 & LD2 \\ 1.28 & 0.76 \\ 4.33 & 1.83 \\ -8.82 & 5.35 \\ -6.00 & 2.26 \end{pmatrix}$	$\begin{pmatrix} LD1 & LD2 \\ 0.72 & 0.97 \\ 6.50 & 0.92 \\ -0.44 & 0.94 \\ -1.27 & -0.78 \end{pmatrix}$	$\begin{pmatrix} LD1 & LD2 \\ -1.08 & 0.61 \\ 5.64 & -0.76 \\ -0.43 & -4.37 \\ -9.66 & -9.21 \end{pmatrix}$	$\begin{pmatrix} LD1 & LD2 \\ -0.34 & 1.68 \\ 4.59 & -1.43 \\ 0.87 & 0.77 \\ -2.38 & -6.50 \end{pmatrix}$
Confusion Matrix	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.97 & 0.03 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.97 & 0.03 \\ 0.00 & 0.03 & 0.97 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.97 & 0.03 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.97 & 0.03 \\ 0.00 & 0.10 & 0.90 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$
AER	**0.01**	**0.02**	**0.01**	**0.04**	**0.00**	**0.00**

It is observed that the classification of the given training data vary based on the procedures. It is noted that the proposed Best Points Selection procedure under MVE and MCD classified the given observations more exactly than the other procedures. The confusion matrix and apparent error rate were computed based on the Discriminant function, which was generated under various procedures is then used to classify the remaining data, which were treated as validation data. The results are summarized in the table 2.

Table 2: Results of Discrimination Analysis under the Various Procedures (Validation Data)

Methods	MLE	MVE	MCD	FSA	BPS_MVE	BPS_MCD
Confusion Matrix	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.95 & 0.05 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.95 & 0.05 \\ 0.00 & 0.05 & 0.95 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.95 & 0.05 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.35 & 0.65 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 0.97 & 0.03 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{pmatrix}$
AER	**0.02**	**0.03**	**0.02**	**0.17**	**0.01**	**0.00**

The computed discriminant function for the training data is validated through the validation data. Almost all the procedures produced more apparent error rate for the validated data set, while using the discrimination function, which was computed by the training data that is, misclassification probabilities are more than that training data. But, the proposed procedure classified exactly under MCD and with little error rate under MVE. Hence, it is concluded that the Best Points Selection (BPS) procedure is produces reliable estimates of the mean vector and covariance matrix for the given multivariate data.

CONCLUSIONS

This paper suggests a robust estimator to estimate the mean vector and covariance matrix of the given multivariate data. It is demonstrated that the established BPS procedure gives the reliable results and more efficient than the other procedures with the help of real data. It is concluded that, the proposed BPS procedure will be applicable in all multivariate techniques, wherever the mean vector and covariance matrix are to be estimated/used and specifically research in high dimensional data analysis.

REFERENCES

1. Anderson, T.W. An Introduction to Multivariate Statistical Analysis. *New York: Wiley*, (1984).
2. Croux.C and Dehon.C. Robust linear discriminant analysis using S-estimators, *The Canadian Journal of Statistics*, Vol. **29**. 473-492, (2001).

3. Fung, W.K. Some diagnostic measures in discriminant analysis, *Statistics and Probability Letters*, Vol. **13**, 279-285, (1992).
4. Hawkins D.M. The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data, Computational Statistics and Data Analysis, 17, 197-210, (1994b).
5. Hawkins D.M. McLachlan G.J. High breakdown linear discriminant analysis, J*ournal of the American Statistical Association*, Vol.**92**, 136-143, (1997).
6. He.X and Fung W.K. High breakdown estimation for multiple populations with applications to discriminant analysis, *Journal of Multivariate Analysis*, Vol. **72**. 151-162, (2000).
7. Hubert M. Van Driseen K. Fast and robust discriminant analysis, *Computat. Statist Data Anal*, vol. **45**. 301-320, (2004).
8. Hubert, M., Rousseeuw, P. J. and Van Aelst, S. High-Breakdown Robust Multivariate Methods, *Statistical Science*, 23, 92-119, (2008).
9. Indrani Basak. Robust M-estimation in Discriminant Analysis. *The Indian Journal of Statistics*, Vol.**60**, Series B, Pt. 2, PP.246-268, (1998).
10. Maronna, R.A. Robust M-estimators of multivariate location and scatter, *Annals of Statistics*, Vol.**4**, 51-67, (1976).
11. Muthukrishnan.R and K.Mahesh. Performance of Classical Robust Linear Discriminant Analysis, *International Journal Statistics and Analysis* ISSN 2248-9959 Vol.2, Number 3, PP.239-243, (2012).
12. Muthukrishnan, R., and K. Mahesh and M. Radha. Robust Methods for Regression Outliers using R, *International Journal of Statistics and Analysis*, ISSN 2248-9959 Vol.2,Number 4, PP-411-146, (2012).
13. Muthukrishnan.R, and K.Mahesh. Evaluation of classical and robust discriminant methods under apparent error rate. *International journal of Current Research*, ISSN: 0975-833X- Vol.5, Issue.10, PP.2817-2820, (2013).
14. Rousseeuw, P.J. Multivariate estimation with high breakdown point, Mathematical Statistics and Applications, Vol. B (Grossmann et al., eds.), 283-297, (1985).
15. Rousseeuw, P.J. and B.Van Zomeren. Unmasking multivariate outliers and leverage points, *Jour. Amer. Statist. Assoc.* Vol. **85** 633-651, (1990).
16. Rousseeuw,P.J. and Van Driessen.K. A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, Vol.**41**. 212-223, (1999).
17. Narayan C.Giri Multivariate Statistical Analysis. Second edition, Marcel Dekker,Inc. Newyork, (2004).
18. Todorov.V, Neykov N. Neytchev. Robust two group discrimination by bounded influence regression, *Computat.Statist Data Anal*, Vol. **17**. 289-302, (1994).
19. Valentin Todorov and Ana M.Pires. Comparative Performance of Several Robust Linear Discriminant Analysis methods. REVSTAT- Statistical Journal 5(1): 63-83, (2007).